

# Using the Grid to Analyse Linked Datasets

The GRID for Research  
ESRC Research Methods Festival  
July 18, 2006

# Outline

- Background
  - **Grid-enabled micro-econometric data analysis,(GEMEDA),  
an e-Social Science pilot demonstrator project**
- Data
- Empirical Analysis
- Implementation
- Results Presentation
- An Empirical Worker's Perspective

# Background

## **Social Science Problem** *And Policy Issue*

- Researchers frequently have to use more than one data set in order to obtain a more complete answer to their questions

*E.g. What do we know about ethnic minority economic welfare when it is disaggregated by group and geography*

- One data set may provide a large sample of the target population, but offer incomplete coverage of the topics of interest

*Census data can lack direct measures of income*

- Another data set with coverage of the topics of interest may not sample the target population adequately

*Survey data yield minority samples that may be too small for meaningful results to be obtained*

- The econometric analysis involved belongs in the group of data linkage methods. The case depends on how the data sets can be formally matched.
- As there are no identifiable common units in the problem considered here, it is one of statistical data fusion.
- The actual methodology used is taken from the poverty mapping literature.
- The survey data can be viewed as the donor data set, with the census based data being the recipient data set.

# Data

- The British Household Panel Survey (BHPS) provides the small scale survey data.
  - BHPS is a longitudinal (panel) study with yearly waves.
- The Sample of Anonymised Records (SARs) provides the large scale Census data.
  - SARs are a random sample of individuals and households from the UK Census
- Uses 1991 data because of projected confidentiality restrictions on the publicly available version of the 2001 SARs.
  - 2% sample of individuals, 1% sample of households.

## **Data Issues**

- It is time consuming dealing with different data sets when they are available in a wide variety of user unfriendly formats.
- Need common and coherent variable definitions for the donor and recipient data sets.

## **Addressed by the Data Grid?**

- The issues suggest a workflow which becomes messy when dealing with communication between the steps in the overall analysis: data extraction, computation, and results presentation.
- This is alleviated by hosting the data on a data grid.

# Empirical Analysis

1. estimate a statistical model using the BHPS data
  - taking account the heterogeneous nature of the survey data
2. use the results to provide income predictions for the SARs data
  - uses parameter estimates from 1.
3. use the income predictions along with other results to estimate poverty measures and their standard errors.
  - headcount (% below a given poverty line)
  - poverty gap (% distance from a poverty line)
4. present these poverty measures
  - by UK regions, SARs areas, GB profiled areas
  - for ethnic groups (by gender if using individuals)

## **Empirical Issues**

- Statistical inference may be difficult in combined data problems. Underlying theoretical assumptions may be too strong. Calculation may be difficult.
- Simulation techniques such as statistical bootstrapping may address these difficulties, however, these can be computational intensive.

## **Addressed by the Computational Grid?**

- Yes. These techniques are embarrassingly parallizable and well suited to implementation on High Performance Computers (HPC).

# Implementation Issues

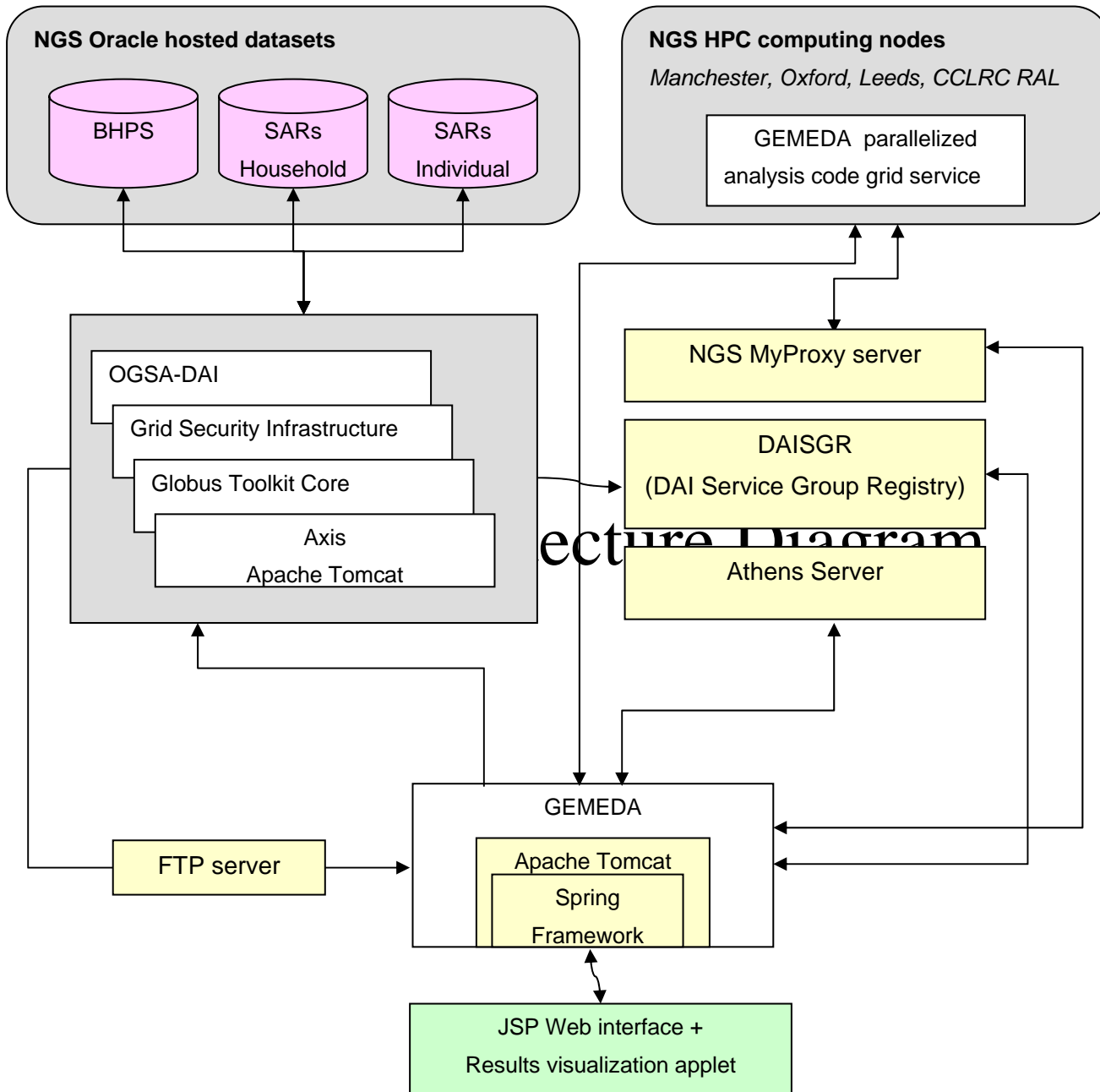
1. Social scientists rarely have easy access to HPC or HTC resources: **staff, human capital, equipment**
2. Need to avoid deploying complex middleware stacks on end-user's computers.

## Suggested the following Grid solution

- Tap into established trends & ongoing investments in UK e-Science
  - National Grid Service (NGS), solves 1.
  - Web service technologies (portals), solves 2.

# The NGS Implementation

- Grid has four core clusters.
  - Two specialising in data hosting
  - Two in compute intensive operations.
- Oracle databases.
  - Accessed via OGSA-DAI
- Parallelization using MPI.
  - Available for more languages and systems than OpenMP
- Globus toolkit.
  - GridFTP.
- MyProxy server for e-certificates.
  - single sign on



# Presentation of Results

- GIS style choropleth map,
  - *area colouring represents range of poverty measure*
- with linked plot
  - *boxplot style graphic of income for main category of interest for a chosen area.*
- Requires mapping data
  - *from EDINA, Athens authenticated*
- Implemented as a Java applet
  - *uses opensource GeoTools java library 2.*

GEMEDA map - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

https://pascal.mvc.mcc.ac.uk:8443/gemeda/gemedaMap/map.jsp?id=1138296800988

Google Met Office

[ Met2\_pov = 0 ]  
 [ Met2\_pov > 0.0 ] AND [ Met2\_pov <= 10.0 ]  
 [ Met2\_pov > 10.0 ] AND [ Met2\_pov <= 20.0 ]  
 [ Met2\_pov > 20.0 ] AND [ Met2\_pov <= 25.0 ]  
 [ Met2\_pov > 25.0 ] AND [ Met2\_pov <= 33.3 ]  
 [ Met2\_pov > 33.3 ] AND [ Met2\_pov <= 50.0 ]  
 [ Met2\_pov > 50.0 ] AND [ Met2\_pov <= 66.6 ]  
 [ Met2\_pov > 66.6 ] AND [ Met2\_pov <= 100.0 ]

Region/SARs Area

White  
 Black Caribbean  
 Black African  
 Black other  
 Indian  
 Pakistani  
 Bangladeshi  
 Chinese  
 Other-asian  
 Other-other

Male  
 Female  
 All

*gender buttons*

*area toggle*

*ethnic group buttons*

### UK Male Imputed Income

Ethnic Group	Monthly Income (Approximate)
White	500 - 2500
Black Caribbean	500 - 2500
Black African	500 - 2500
Black other	500 - 2500
Indian	500 - 2500
Pakistani	500 - 2500
Bangladeshi	500 - 2500
Chinese	500 - 2500
Other-asian	500 - 2500
Other-other	500 - 2500

Monthly income

show data

Applet MapViewer started

pascal.mvc.mcc.ac.uk:8443

GEMEDA map - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

https://pascal.mvc.mcc.ac.uk:8443/gemeda/gemedaMap/map.jsp?id=1138296800988

Google Met Office

[ Met2\_pov = 0 ]  
 [ Met2\_pov > 0.0 ] AND [ Met2\_pov <= 10.0 ]  
 [ Met2\_pov > 10.0 ] AND [ Met2\_pov <= 20.0 ]  
 [ Met2\_pov > 20.0 ] AND [ Met2\_pov <= 25.0 ]  
 [ Met2\_pov > 25.0 ] AND [ Met2\_pov <= 33.3 ]  
 [ Met2\_pov > 33.3 ] AND [ Met2\_pov <= 50.0 ]  
 [ Met2\_pov > 50.0 ] AND [ Met2\_pov <= 66.6 ]  
 [ Met2\_pov > 66.6 ] AND [ Met2\_pov <= 100.0 ]

Region/SARs Area

- White
- Black Caribbean
- Black African
- Black other
- Indian
- Pakistani
- Bangladeshi
- Chinese
- Other-asian
- Other-other

- Male
- Female
- All

**Northampton Male Imputed Income**

Ethnic Group	Min	Q1	Median	Q3	Max
White	~200	~400	~700	~1100	~1800
Black Caribbean	~100	~300	~600	~1000	~1500
Black African	~100	~100	~100	~100	~100
Black other	~100	~100	~100	~100	~100
Indian	~100	~400	~600	~1100	~1800
Pakistani	~100	~100	~100	~100	~100
Bangladeshi	~100	~100	~100	~100	~100
Chinese	~100	~100	~100	~100	~100
Other-asian	~100	~100	~100	~100	~100
Other-other	~100	~100	~100	~100	~100

475329°54.2'E 261385°36.8'N

show data

Applet MapViewer started

pascal.mvc.mcc.ac.uk:8443

# An Empirical Worker's Perspective

- **What has the NGS' Grid solution provided?**
  - ✓ Equipment.
    - middle range HPCs (and software) are accessible to Social Scientists.
  - ✓ Staff.
    - maintenance, development and implementation of middleware.
  - ✓ Human capital.
    - knowledge and experience gained by researchers stays relevant and remains within the Social Sciences

- **What are the weaknesses of the present implementation?**

- **Sustainability.**

- staff, maintenance, ...

- re-engineering (Globus WSRF, OGSA-DAI & SQL Server, ...)

- **Reliability.**

- error reporting for distributed architecture, documentation

- robustness of compute nodes, interoperability of middleware

- **Security.**

- Athens authentication, e-certificates, ...

- firewalls, ...

- **Modeling.**

- more variables, more data sets, repeat data sets (2001)

- interactivity, visualization, ...

- **Grid Limitations**

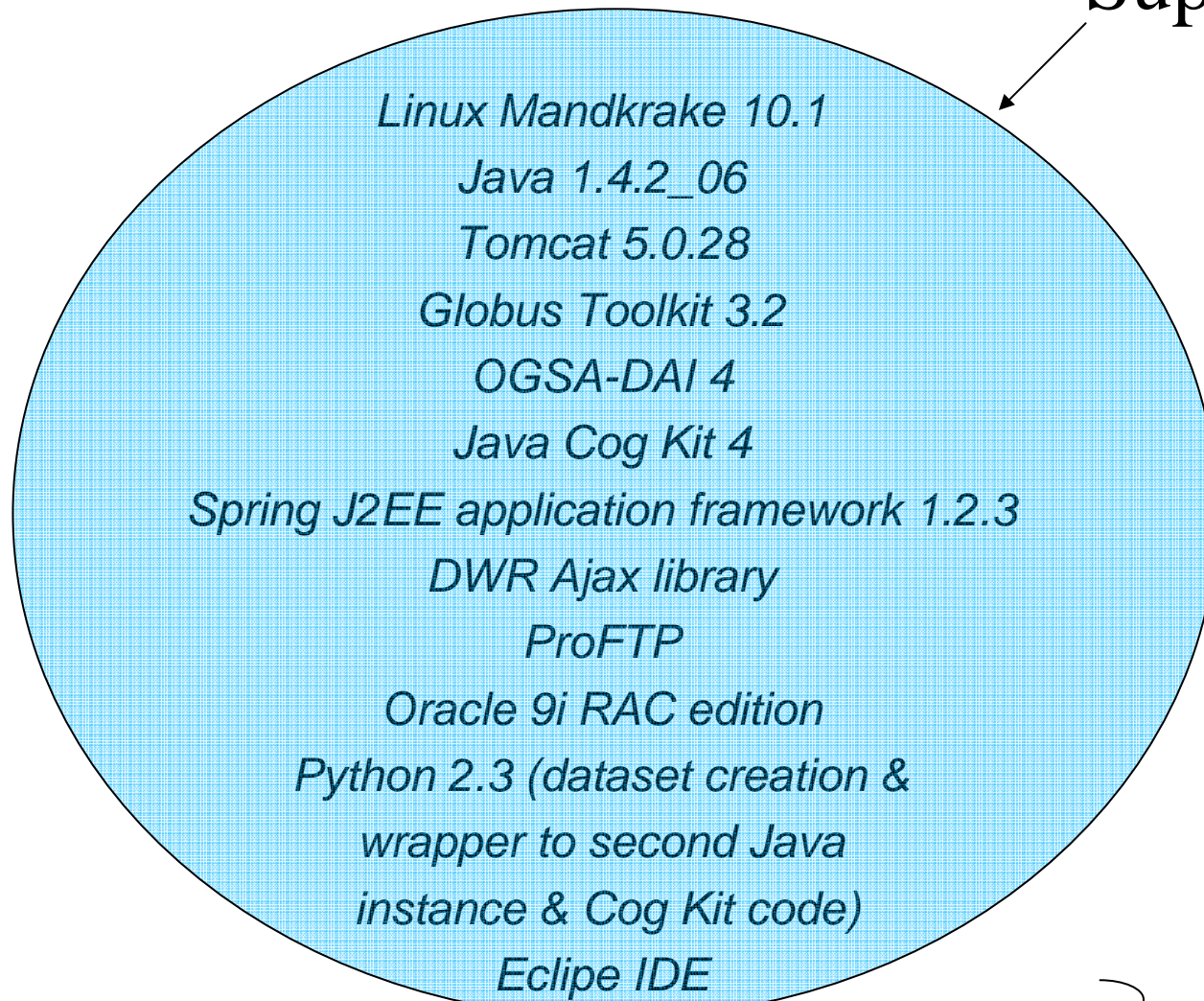
- **Having grid-enabled data does not mean you have the process of data cleaning or data manipulation grid-enabled.**
- **Computational grids require compute resource brokerage (NGS is batch).**
- **Supporting code much larger than business code, specialist elements of the project were easiest to develop.**

- **General Limitations**

- **Data disclosure controls, confidentiality restrictions and proprietorship issues.**
- **Expertise of researchers within discipline areas.**

# Software Used:

Supporting



*Fortran95+MPI*

*C (utiliies for converting output for use with the map)*

*GeoTools (Java library for mapping)*

Specialist

## Project Team

Simon Peters, Ken Clark SoSS (economics)

Pascal Ekin, Anja Le Blanc, Stephen Pickles Manchester  
Computing

Project portal and service: <http://pascal.mvc.mcc.ac.uk/gemeda/>

## Acknowledgements

Celia Russell, Mike Jones

SAMD

Mark Birkin, Andy Turner

Hydra I Grid (now MoSeS)

Keith Cole

ConvertGrid

Matt Ford

NGS

